Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v7n2p432

A simple and conservative empirical likelihood function

By Miller G.

Published: 14 October 2014

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, 2014, 432-437
DOI: 10.1285/i20705948v7n2p432

A simple and conservative empirical likelihood function

Guthrie Miller*

Santa Fe, New Mexico, USA

Published: 14 October 2014

Multiple measurements with an unknown Gaussian likelihood function are treated probabilistically. The likelihood function $L(\mu) \propto \left((n-1)s_x^2 + n(\mu - \overline{x})^2\right)^{-n/4}$ is derived where μ is the true value, $\overline{x}$ is the mean, and s_x^2 is the variance obtained from the n measurements.

1 Introduction

As motivation for this little calculation, imagine the following specific situation. You are an Industrial Hygiene professional responsible for the health and safety of workers at your plant. You have occupational exposure data (for example, exposure to lead aerosol) consisting of three measurements of 0.2, 0.05, and 0.1, while the occupational exposure limit is 1. So, you need to calculate the probability distribution of the true value of the exposure, given the measurement results obtained. You are particularly interested in the probability that the true value of the exposure exceeds l.

This application involves protecting the health of the worker, so a very conservative approach is warranted.

Using the traditional (non Bayesian) approach with a lognormal model, one would start by calculating the geometrical mean as 0.1 and the geometric standard deviation as 2 by taking the exponentials of the average and standard deviations of the logs of the measurement values. Then an upper confidence limit would be calculated as the geometric mean times the geometric standard deviation raised to some power like 1.645 or 2. However one then considers the upper confidence limit of the geometric standard deviation derived from just three data points, which leads to a much higher upper limit.

*Corresponding author: guthriemiller@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

The probabilistic (Bayesian) approach is the logical one, and it is conceptually simplest and most straightforward. It produces at the end what is desired but never provided with the traditional approach, namely a plot showing the probability distribution for the quantity of interest. It *is* necessary to provide prior probability distributions for all the parameters, but this is natural, because the logic demands it.

Elements of the analysis presented here are similar to material in Chapter 8 of the book by Sivia (Sivia and Skilling, 2006), but the main result quoted in the abstract seems not to have been stated before.

The likelihood function derived here, which assumes a logarithmic prior on the true standard deviation, can serve as a most conservative limiting case for many data analysis situations.

2 The Calculation

For the more general version of this problem the data consist of some number n of measurements giving results x_i for $i = 1, \ldots n$. The probability distribution of each x_i is assumed to have a Gaussian form

$$P(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \quad , \tag{1}$$

with unknown mean μ and standard deviation σ. To go over to a lognormal, $\mu \to \log(\mu)$.

The parameters are μ and σ . The likelihood function for μ and σ is the product of $P(x_i|\mu,\sigma)$ for $i = 1, \ldots n$, and therefore

$$L(\mu,\sigma) \propto \frac{1}{\sigma^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = \frac{1}{\sigma^{n/2}} \exp\left(-\frac{X}{2\sigma^2}\right) \quad , \tag{2}$$

with

$$X \equiv (n-1)s_x^2 + n(\mu-\overline{x})^2 \quad , \tag{3}$$

where, just by doing some algebra, one sees that the data enter in via X only through the sample mean and standard deviation defined by

$$\begin{aligned} \overline{x} &\equiv \frac{1}{n}\sum_{i=1}^{n} x_i \\ s_x^2 &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \quad . \end{aligned} \tag{4}$$

One needs to assume some definite form for the priors on μ and σ . We will assume an alpha prior (Miller, 2013) on μ representing the prior probability distribution on true amount μ (not $\log(\mu)$)

$$P(\mu) \propto \mu^{\alpha-1} \quad , \tag{5}$$

where α is a number less than or equal to 1, with 1 giving a flat prior and small values giving a concentration at $\mu = 0$. This form is assumed to extend from a minimum value $\mu_{\min}$, which can be 0, to a maximum value $\mu_{\max}$, which is unimportant as long as it is greater than the largest conceivable amounts for the problem (so that the data provide the upper cutoff on the posterior probability).

For σ, the natural prior is logarithmic by scale invariance, but for more generality

$$P(\sigma) \propto \sigma^{-p} \quad , \tag{6}$$

where $p = 1$ gives $P(\sigma)d\sigma \propto d(\log \sigma) = d\sigma/\sigma$. This form extends from some small minimum value $\sigma_{\min}$, which can be 0, to a maximum value $\sigma_{\max}$, which can be infinite.

Under these assumptions using the elementary rules for conditional probability (Bayes theorem), the probability distribution of true amount, given the measurement results, can be immediately written down:

$$P(\mu' < \mu | data) = \frac{\int_{\mu_{\min}}^{\mu} \left(\int_{\sigma_{\min}}^{\sigma_{\max}} L(\mu', \sigma) P(\sigma) d\sigma \right) P(\mu') d\mu'}{\int_{\mu_{\min}}^{\mu_{\max}} \left(\int_{\sigma_{\min}}^{\sigma_{\max}} L(\mu', \sigma) P(\sigma) d\sigma \right) P(\mu') d\mu'} \quad . \tag{7}$$

Results are shown in Fig. 1 for the example problem. The solid curves show the posterior cumulative probability (probability that μ is less than the x-axis value) for a uniform prior on μ ($\alpha = 1$) and a logarithmic prior on σ ($p = 1$). The dashed curves, which are shifted to smaller values relative to the solid curves, are for the prior on μ concentrated at 0 ($\alpha = 0.01$). In this figure, the prior on σ has $\sigma_{\max} = 3$, corresponding to considering GSD's greater than 20 to be unreasonable and not allowing them.

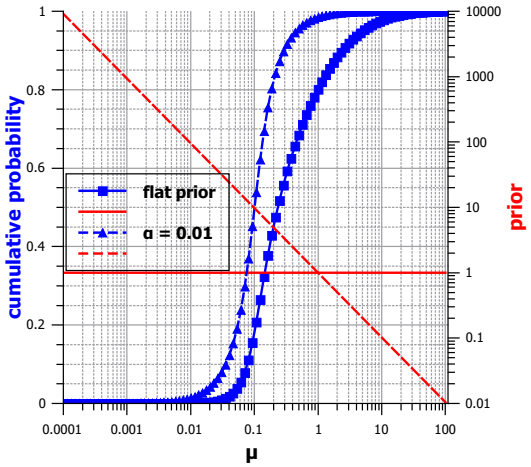

Figure 1: Cumulative probability of true amount μ given three measurements of 0.2, 0.05, and 0.1 with the priors as discussed. The prior that is concentrated at 0 shifts the probability curve down to smaller values of μ .

The priors on μ and σ would need to be justified empirically. For similar situations, what is the distribution of μ actually observed? If only small values are seen, then a prior with α small would be appropriate. Similarly, what are the values of s_x actually observed?

In the limit $\sigma_{\min} \to 0$ and $\sigma_{\max} \to \infty$, the integral over $d\sigma$ converges and gives, for a lognormal, the result

$$L(\mu) \propto X^{-n/4-(p-1)/2} \quad , \tag{8}$$

with X defined by Eq. 3, but this is a very broad distribution with a flat prior unless n is 10 or so, as shown in Fig. 2. This would seem to say that the number of measurements should always be about this number or greater.

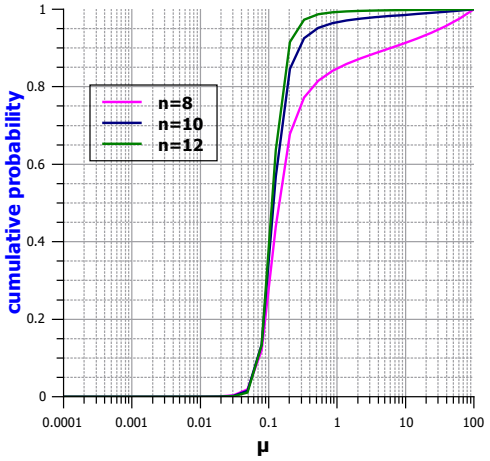

Figure 2: Flat-prior cumulative probability of true amount μ for different numbers of measurements.

Returning to the specific example problem, one would conclude using the probabilistic approach that more data are needed in order to be sure that the exposure is less than 1, either old data for similar situations assembled together to justify the α prior, or new data to use with a flat prior.

3 Algebraic Details

$$\int_{\sigma_{\min}}^{\sigma_{\max}} P(\sigma) L(\mu, \sigma) d\sigma = \int_{\sigma_{\min}}^{\sigma_{\max}} \frac{1}{\sigma^{n/2+p}} \exp\left(-\frac{X}{2\sigma^2}\right) d\sigma \quad . \tag{9}$$

Change variables to

$$\begin{aligned} Z &= \frac{X}{2\sigma^2} \\ dZ &= -\frac{X}{\sigma^3} d\sigma \end{aligned} \tag{10}$$

obtaining

$$\begin{aligned} \int_{\sigma_{\min}}^{\sigma_{\max}} P(\sigma) L(\mu, \sigma) d\sigma &= \int_{Z_{\min}}^{Z_{\max}} \frac{\sigma^{3-p-n/2}}{X} \exp(-Z) dZ \\ &= \frac{2^{n/4+(p-1)/2-1}}{X^{n/4+(p-1)/2}} \int_{Z_{\min}}^{Z_{\max}} Z^{n/4+(p-1)/2-1} \exp(-Z) dZ \quad . \end{aligned} \tag{11}$$

The upper limit on the right can be taken to $Z_{\max} \to \infty$, because the exponential $\exp(-Z)$ guarantees convergence. The lower limit can be taken to $Z_{\min} = 0$, because the integral,

$$\int_0^{Z_{\min}} Z^{n/4+(p-1)/2-1} \exp(-Z) dZ = \frac{Z_{\min}^{n/4+(p-1)/2}}{n/4 + (p-1)/2} \quad . \tag{12}$$

converges when $n/4 + (p-1)/2 > 0$. For a non-zero $Z_{\min}$ (finite $\sigma_{\max}$),

$$\begin{aligned} L(\mu) &\propto \int_0^{\sigma_{\max}} P(\sigma) L(\mu, \sigma) d\sigma \\ &\propto X^{-n/4-(p-1)/2} \int_{\frac{X}{2\sigma_{\max}^2}}^{\infty} Z^{n/4+(p-1)/2-1} \exp(-Z) dZ \quad . \end{aligned} \tag{13}$$

The quantity given in Eq. 13, the likelihood function, when multiplied by the prior on μ gives the probability distribution of μ that is sought. The likelihood function is plotted in Fig. 3 for the example problem. For very large n, the likelihood function approaches a lognormal with logarithmic standard deviation $S = s_x \sqrt{2/n}$.

Acknowledgement

This example problem appears in a current workbook put out by the National Industrial Hygiene Association to prepare candidates for their certification examination, in the particular section written by Dr. Paul Hewett entitled "Introduction to Bayesian Decision Analysis". The author also acknowledges Prof. Devinder Sivia for a helpful exchange of emails.

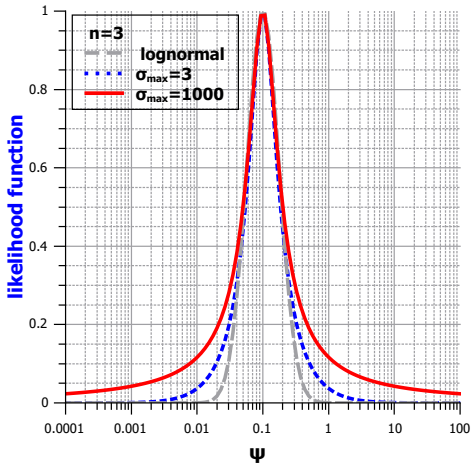

Figure 3: Likelihood function from Eq. 13 for example problem with different limits on prior on σ and showing the lognormal approximation valid for large n.

References

Miller, G. (2013). *Probabilistic Interpretation of Data–A Physicist's Approach.* Lulu Publications.

Sivia, D. S. and Skilling, J. (2006). *Data Analysis–A Bayesian Tutorial, 2nd Ed.* Oxford Science Publications.